An analysis of psychoacoustically-inspired matching pursuit decompositions of speech signals

Khalid Daoudi[1], *Nicolas Vinuesa*[2]

[1]INRIA Bordeaux-Sud Ouest, GeoStat team. France
[2]GIN UMR5296, CNRS-Université de Bordeaux. France
khalid.daoudi@inria.fr, vinuesa.nico@gmail.com

Abstract

Matching pursuit (MP), particularly using the Gammatones dictionary, has become a popular tool in sparse representations of speech/audio signals. The classical MP algorithm does not however take into account psychoacoustical aspects of the auditory system. Recently two algorithms, called PAMP and PMP have been introduced in order to select only perceptually relevant atoms during MP decomposition. In this paper we compare this two algorithms on few speech sentences. The results suggest that PMP, which also has the strong advantage of including an implicit stop criterion, always outperforms PAMP as well as classical MP. We then raise the question of whether the Gammatones dictionary is the best choice when using PMP. We thus compare it to the popular Gabor and damped-Sinusoids dictionaries. The results suggest that Gammatones always outperform damped-Sinusoids, and that Gabor yield better reconstruction quality but with higher atoms rate.

Index Terms: Matching pursuit, Time-frequency decomposition, Sparse representation, Gammatones, Perceptual models.

1. Introduction

During the last two decades, the Matching pursuit (MP) algorithm [1] has been widely used as a powerful tool for spare representation of signals using redundant dictionaries of time-frequency functions (atoms). MP is a greedy algorithm which iteratively approximates a signal $x(t)$ by a projecting it onto an overcomplete dictionary D of atoms ϕ_θ:

$$R_x^m(t) = \langle R_x^m(t), \phi_\theta \rangle \phi_\theta + R_x^{m+1}(t), \tag{1}$$

with $R_x^0(t) = x(t)$ at the first iteration $m = 0$. At each iteration m, a single atom ϕ_m is selected such that:

$$\phi_m = \arg\max_{\phi_\theta \in D} |\langle R_x^m(t), \phi_\theta \rangle| \tag{2}$$

where $\langle \cdot, \cdot \rangle$ is (generally) the Hermitian inner product. The signal $x(t)$ can be thus decomposed as:

$$x(t) = \sum_{m=1}^{M} \sum_{i=1}^{n_m} s_i^m \phi_m(t - \tau_i^m) + \epsilon(t), \tag{3}$$

where τ_i^m and s_i^m are the temporal position and weight of the ith instance of the atom ϕ_m, respectively. The notation n_m indicates the number of instances of ϕ_m, which need not to be the same across atoms, and M indicates the number of different atoms.

Recently a toolkit which efficiently implements the classical matching pursuit algorithm has been released: the Matching Pursuit ToolKit (MPTK) which is based on the work in [2] and can be downloaded from http://mptk.irisa.fr. It can be installed on various platforms (Windows, Linux and Mac OSX) and is now massively used as it is the best available toolkit for (classical) MP analysis. MPTK provides fast implementation of different kind of dictionaries, including the Gabor dictionary.

In the field of speech/audio coding, it has been argued in [3, 4] that a relatively small dictionary of Gammatone atoms allow efficient coding of natural sounds using MP. The motivation behind this work is that early psychoacoustic experiments used Gammatone functions as a model of basilar membrane displacement [5] and where found to approximate cochlear responses of the cat [6]. Later it was stated in [7] that Gammatone functions also delineate the impulse response of human auditory filters. Real-valued Gammatone filters can be seen as gamma-modulated sinusoids and are defined as:

$$\gamma(t) = t^{n-1} e^{-2\pi b ERB(f_c)} cos(2\pi f_c t), \tag{4}$$

where f_c is the central frequency distributed on ERB (equivalent rectangular bandwidth) scales, n is the filter order, and b is a parameter that controls the bandwidth of the filter. $ERB(f_c) = 24.7 + 0.108 f_c$, $n = 4$ and $b = 1.019$ are commonly used as parameters.

On the other hand, classical MP (used in [3, 4]) focus on minimizing the energy of residuals at each iteration. Thus, it does not take into account psychoacoustical aspects of the auditory system which are crucial in any codec development. To address this issue, a psychoacoustically-adaptive inner product, considering frequency masking effects in sinusoidal decompositions, was presented in [8] and later refined with a perceptual model in [9]. The resulting algorithm is called PAMP. More recently a new perceptual model, called PMP, was introduced in [10], taking into account both temporal and frequency masking effects. Similar to the perceptual model embedded in MPEG coders, the goal of psychoacoustically-based MP algorithms is to discard the perceptually irrelevant structures of the input signal and therefore increase the coding efficiency.

In this article we first experimentally compare these two algorithms when using a dictionary of Gammatone atoms. Our experiments suggest that PMP, which also has the strong advantage of including an implicit stop criterion, always outperforms PAMP as well as classical MP. We then raise the question of whether the Gammatones dictionary is the best choice when using PMP. We thus compare it to the popular Gabor and damped-sinusoids dictionaries. The results suggest that Gammatones always outperform damped-sinusoids, and that Gabor yield better reconstruction quality but with higher atoms rate.

The paper is organized as the following. In section 2, the two psychoacoustically-based MP algorithms are briefly described. Section 3 presents and analyzes the results of com-

parison between classical MP, PAMP and PMP. In section 4 we present an evaluation of PMP when using different dictionaries. Finally, we draw our conclusion and perspectives in section 5.

2. Psychoacoustically-inspired matching pursuits

In this section we give a short description of PAMP and PMP. Details about these algorithms can be found in [8, 9] and [11], respectively.

2.1. PAMP

PAMP [8, 9] relies on a perceptual model which predicts masked thresholds for sinusoidal distortions in audio coding. This model exploits the simultaneous masking effect (frequency masking) in order to determine what distortion level can be allowed such that it is perceptually not detectable. The model is based on a perceptual distortion measure [12] which estimates the probability that subjects can detect a distortion signal in the presence of a masking signal. This distortion measure defines a norm:

$$\|x\|^2 = \sum_f \hat{\alpha}(f)|\hat{x}(f)|^2 \tag{5}$$

where $\hat{x}(f)$ is the Fourier transform of the signal x and $\hat{\alpha}(f)$ is a real and positive weighting function representing the inverse of the masking curve for sinusoidal distortions. The norm is induced by the inner product:

$$\langle x, y\rangle = \sum_f \hat{\alpha}(f)\hat{x}(f)\hat{y^*}(f), \tag{6}$$

This inner product is then used in Eq. 2 instead of the classical Hermitian inner product in order to select the sinusoidal components of the signal in a MP decomposition with a dictionary of sinusoids. At the first iteration, i.e., when the residual equals the signal, $\hat{\alpha}(f)$ is set as the inverse of the threshold-in-quiet. Then, at each iteration $\hat{\alpha}(f)$ takes into account the sinusoidal distortion caused by the atom selected at the previous iteration.

2.2. PMP

PMP [11] relies on a perceptual model which take into account both temporal and frequency masking effects (as opposed to PAMP which considers frequency masking only). In PMP, a dictionary of Gammatone atoms is used and a masking pattern is created (and progressively updated) to determine a masking threshold at all time indexes and atom central frequencies.

At the first iteration, the masking pattern is set to the threshold-in-quiet, as in PAMP, for all time indexes. Then, all the inner products in Eq. 2 which are below the masking pattern are set to zero, meaning that projections that are below he threshold-in-quiet are ignored. Once the first atom has been selected, which will act as a masker, the masking pattern is elevated in a time interval around the atom temporal position and in the two adjacent critical bands. The updated masking pattern is then again used as a threshold, setting to zero all the inner products below it, thereby avoiding the search of atoms that would be masked by previously selected ones, i.e. perceptually irrelevant. This process is repeated until no inner product is above the masking threshold, meaning that there is no audible part left in the residual, and the algorithm stops. This implicit and perceptually-motivated stop criterion is a strong advantage over classical MP and PMP.

3. Comparison between MP, PMP and PAMP

In this section, we experimentally compare the performance of the two psychoacoustically-based and the the classical matching pursuit algorithms. Since Gammatones have become popular waveforms in sparse speech/audio representations, we perform this comparison in the setting of MP using Gammatones dictionaries. The main idea is to analyse the behavior of MP when (Gammatone) atom selection is performed by the two perceptual models. We recall however that, while PMP has been introduced in this setting, PAMP have been developed in the framework of sinusoidal decompositions. By doing such a comparison, we are thus evaluating the behavior of PAMP when distortions are generated by Gammatone components.

We use four sentences from the TIMIT database [13] for the experiments. We selected these excerpts such that both speaker and phonetic variability is achieved: two male (sx54 and sx221) and two female (sx23 and sx136) speakers from different geographic regions are used in this study. The following speakers were used: mbma1 (sx54), fdrw0 (sx23), fgcs0 (sx136) and mcre0 (sx221). All files are sampled at 16 kHz with 16-bit quantization.

While signal to noise ratio (SNR) is a valid measure for *waveform* reconstructability, for audio coding problems this does not necessary reflect the perceived quality of the reconstruction. Therefore we use the well-known perceptual quality assessment measure PESQ [14] to estimate mean opinion scores (MOS). PESQ gives a continuous grading scale from 1 (very annoying) to 5 (no perceptual difference between original and reconstruction).

Since PMP is the only algorithm which implicitly has a perceptual stop criterion, we use the latter as an operating point to compare the 3 algorithms. We first run PMP and then compute the atoms rate per sample when it stops. Then, MP and PAMP are stopped when they reach this atoms rate. The experimental results of this process are are shown in Table 1, for Gammatones dictionaries with 32, 64 and 128 atoms. A first observation is that PMP always yield a PESQ value above 3.5. Moreover, our listening tests confirm that the reconstruction quality is good without being perceptually transparent. This shows that the perceptual model of PMP achieves the desired goal. A second observation is that PMP always slightly outperforms MP and significantly outperforms PAMP. Finally, these results suggest that a good choice for the number of Gammatones in the dictionary is 64.

Figure 1 displays the evolution of the 3 algorithms, iteration after iteration, until they they reach the atoms rate given by PMP. The figure corresponds to only one sentence, but the behavior is very similar for the 4 sentences. Because the masking pattern is updated with the masking effect caused by each new atom, PMP behaves exactly like MP until most of the masker atoms have been extracted; only then (around atoms rate of 0.05) the newly selected atoms are perceptually relevant and the difference can be appreciated. From this rate, PMP starts selecting atoms which are perceptually relevant and thus yields higher PESQ values, while MP selects atoms which minimize the residual energy and thus yields higher SNR values. The weak performances of PAMP are most probably due to the fact that distortions generated by Gammatone decompositions do not satisfy the hypothesis made on distortions obtained in sinusoidal modeling.

In Table 2, we provide the atoms rate required by MP and PAMP to achieve the same PESQ value as PMP (at stop-

File	Dictionary	PESQ-PMP	PESQ-MP	PESQ-PAMP	Atoms Rate
sx54					
	32 Gammatones	3.66	3.56	3.17	0.09
	64 Gammatones	3.70	3.61	3.22	0.08
	128 Gammatones	3.70	3.61	3.21	0.08
sx23					
	32 Gammatones	3.77	3.45	3.07	0.15
	64 Gammatones	3.84	3.62	3.08	0.14
	128 Gammatones	3.85	3.67	3.15	0.14
sx136					
	32 Gammatones	3.60	3.43	2.98	0.09
	64 Gammatones	3.64	3.42	3.05	0.08
	128 Gammatones	3.62	3.47	3.08	0.08
sx221					
	32 Gammatones	3.55	3.41	3.19	0.09
	64 Gammatones	3.58	3.47	3.33	0.08
	128 Gammatones	3.59	3.49	3.35	0.08

Table 1: *PESQ and atoms rate using Gammatone dictionary.*

ping atoms rate), for 64 Gammatones. It is clear that PMP exhibits the highest efficiency among the three algorithms, as MP requires up to 40% and PAMP up to 80% more atoms to achieve the same perceived quality. All these experimental results suggest that PMP is a very good algorithm for efficient and perceptually-consistent sparse speech representations and coding.

File	Atoms rate PMP	Atoms rate MP	Atoms rate PAMP
sx54	0.08	0.11	0.13
sx23	0.14	0.17	0.23
sx136	0.08	0.10	0.15
sx221	0.08	0.10	0.11

Table 2: *Atoms rate required by MP and PAMP to reach the same PESQ value as PMP.*

4. Comparison of different dictionaries using PMP

Given the results of the previous section which are in favor of PMP, we now focus on the latter and raise the following question: is the Gammatones dictionary the best choice when using PMP? This question has been indeed central in classical matching pursuit. The original MP algorithm [1] used the Gabor dictionary defined as:

$$g_\theta(t) = K_\theta e^{-\pi\left(\frac{t-\tau}{s}\right)^2} e^{i\omega(t-\tau)}, \tag{7}$$

for index $\theta = \{s, \tau, \omega\}$ where s is the scale, τ the time translation , ω the frequency modulation, and K_θ such that $\|g_\theta\| = 1$.

Probably the most known work on this matter is [15], where the authors argued that a dictionary which consists only of atoms that exhibit symmetric time-domain behavior are not well suited for modeling asymmetric events such as transients in audio signals. They proposed the use of structured overcomplete dictionaries of damped sinusoids (DS) defined as:

$$d_\theta(t) = K_\theta \lambda^{(t-\tau)} e^{i\omega(t-\tau)} u(t-\tau), \tag{8}$$

File	Dictionary	PESQ-PMP	Atoms rate
sx54			
	64 Gammatones	3.69	0.08
	64 Gabor	3.85	0.12
	64 D-Sinusoids	3.55	0.08
sx23			
	64 Gammatones	3.84	0.14
	64 Gabor	4.04	0.20
	64 D-Sinusoids	3.54	0.15
sx136			
	64 Gammatones	3.64	0.08
	64 Gabor	3.87	0.12
	64 D-Sinusoids	3.39	0.08
sx221			
	64 Gammatones	3.58	0.08
	64 Gabor	3.85	0.12
	64 D-Sinusoids	3.34	0.09

Table 3: *PESQ and atoms rate using different dictionaries.*

for index $\theta = \{\lambda, \tau, \omega\}$ where λ is the damping factor, τ the time translation, ω the frequency modulation, K_θ such that $\|d_\theta\| = 1$ and $u(t-\tau)$ being the step function.

They showed, in the context of classical MP, that DS are more suited for modeling signals with transient behavior than symmetric Gabor atoms. More recently, the work in [16] proposed a comparison of Gabor atoms, complex exponentials and "Fonction d'onde Formantique". The authors argued that the Gabor dictionary performs sufficiently well.

This motivates us to analyse the behavior of PMP when using different dictionaries than Gammatones, within the ERB scale. We thus propose in this section a comparison between Gammatones, damped sinusoids and Gabor atoms.

Table 3 shows the results obtained using 64 atoms per dictionary, within the ERB scale. A first observation is that Gammatones and DS stop at almost the same atoms rate, but Gammatones dictionary outperforms DS. The most important observation is that the Gabor dictionary achieves higher PESQ values than the other dictionaries, but the atoms rate is also considerably higher. If we rely on atom rates as a measure of coding

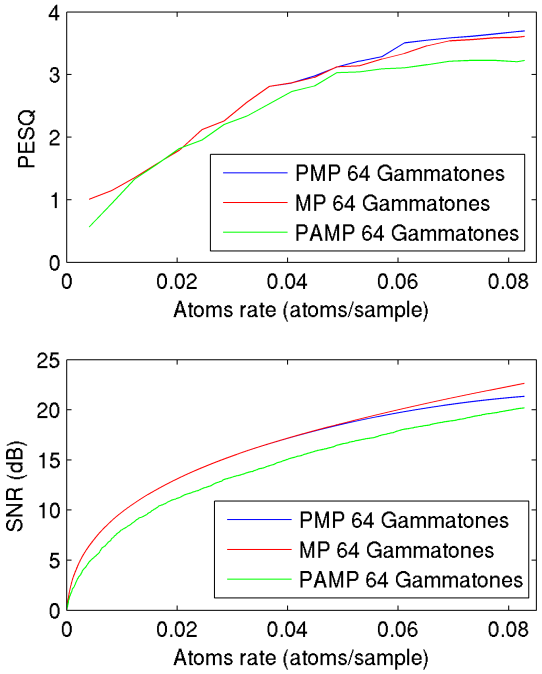

Figure 1: *Comparison of PESQ and SNR vs. atoms rate for sentence sx54 and 64 Gammatones.*

efficiency, we may consider that the Gammatones dictionary is the best choice, given that PMP requires about 50% more Gabor atoms to achieve a relatively smaller gain in PESQ. However, the best way to assess coding efficiency is to use bits per second, and the best way to asses perceptual quality is MOS. Moreover, the ERB scale may not be the optimal choice for Gabor and DS atoms. Finally, we used only 4 sentences in our experiments. A much larger and richer database should be used in order to have a good evaluation. Thus, all these factors (at least) should be taken into account before drawing final conclusions. The question we raised is then still open, but we may still argue that PMP with Gammatones presents a promising potential.

5. Conclusion

In this paper, we first presented an experimental comparison of two psychoacoustically-based matching pursuit algorithms (PMP and PAMP) as well as the classical MP algorithm. The results suggest that PMP always outperforms both MP as PAMP in term of sparsity and perceived reconstruction quality. In a second experiment, we compared different dictionaries (Gammatones, Gabor and damped-sinusoids) using PMP. The results suggest that Gammatones is the best choice if atoms rate is considered as a measure for coding efficiency. All these results suggest that PMP is a very good algorithm for efficient and perceptually-consistent sparse speech representations and coding. However, further work is required in refining the perceptual model in PAMP in order to take into account the distortions generated by Gammatone decompositions more accurately. A more in-depth study of the coding efficiency, using bits per second instead of atoms rate, is also necessary. We believe indeed that latter would allow to mitigate the results of [3, 4]. This will be the purpose of a future work (also using the full TIMIT database in the experiments).

6. Acknowledgments

The authors would like to thank Ramin Pichevar and Hossein Najaf-Zadeh for the help provided in the implementation and fruitful discussions over PMP.

7. References

[1] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *Signal Processing, IEEE Transactions on*, vol. 41, no. 12, pp. 3397–3415, 1993.

[2] S. Krstulovic and R. Gribonval, "MPTK: Matching Pursuit made tractable," in *Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP'06)*, vol. 3, Toulouse, France, May 2006, pp. III–496 – III–499.

[3] E. Smith and M. S. Lewicki, "Efficient coding of time-relative structure using spikes," *Neural Computation*, vol. 17, no. 1, pp. 19–45, 2005.

[4] E. C. Smith and M. S. Lewicki, "Efficient auditory coding," *Nature*, vol. 439, no. 7079, pp. 978–982, 2006.

[5] J. L. Flanagan, "Models for approximating basilar membrane displacement," *The Journal of the Acoustical Society of America*, vol. 32, no. 7, pp. 937–937, 1960.

[6] P. I. Johannesma, "The pre-response stimulus ensemble of neurons in the cochlear nucleus," in *Proceedings of the Symposium of Hearing Theory*, 1972.

[7] R. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," *APU report*, vol. 2341, 1988.

[8] R. Heusdens, R. Vafin, and W. B. Kleijn, "Sinusoidal modeling using psychoacoustic-adaptive matching pursuits," *Signal Processing Letters, IEEE*, vol. 9, no. 8, pp. 262–265, 2002.

[9] S. van de Par, A. Kohlrausch, R. Heusdens, J. Jensen, and S. H. Jensen, "A perceptual model for sinusoidal audio coding based on spectral integration," *EURASIP Journal on Applied Signal Processing*, vol. 2005, pp. 1292–1304, 2005.

[10] H. Najaf-Zadeh, R. Pichevar, L. Thibault, and H. Lahdili, "Perceptual matching pursuit for audio coding," *Audio Engineering Society Convetion, Amsterdam, The Netherlands*, 2008.

[11] R. Pichevar, H. Najaf-Zadeh, L. Thibault, and H. Lahdili, "Auditory-inspired sparse representation of audio signals," *Speech Communication*, vol. 53, no. 5, pp. 643–657, 2011.

[12] S. Van De Par, A. Kohlrausch, G. Charestan, and R. Heusdens, "A new psychoacoustical masking model for audio coding applications," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 2. IEEE, 2002, pp. II–1805.

[13] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "DARPA TIMIT acoustic-phonetic continuous speech corpus," U.S. Dept. of Commerce, NIST, Gaithersburg, MD, Tech. Rep., 1993.

[14] "Perceptual evaluation of speech quality (pesq), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecsITU-T Rec. P.862," 2001.

[15] M. M. Goodwin and M. Vetterli, "Matching pursuit and atomic signal models based on recursive filter banks," *Signal Processing, IEEE Transactions on*, vol. 47, no. 7, pp. 1890–1902, 1999.

[16] B. L. Sturm and J. D. Gibson, "Matching pursuit decompositions of non-noisy speech signals using several dictionaries," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 3. IEEE, 2006, pp. III–III.